CSE574 Introduction to Machine Learning

Gaussian Naive Bayes

Jue Guo

University at Buffalo

March 13, 2024

Outline

Learning Objectives

- Define Bayes' rule and conditional probability
- Define naive Bayes classification
- List and evaluate the assumptions of naive Bayes
- Define Gaussian naive Bayes

Bayes' Rule

- The probability an event occurs may change depending on certain conditions. Ex: Not all emails are equally likely to be spam. An email containing the word "URGENT" is more likely to be spam than an email containing the phrase "Meeting". Conditional probability measures the probability that an event occurs, given another event has also occurred.
- The **conditional probability** of event A, given event B has already occurred, is denoted as $P(A \mid B)$.

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

Bayes' Rule

- In some cases, the known conditional probability is not the condition of interest.
  - Ex: From a random sample of spam emails, $P(\text{ URGENT } \mid \text{ Spam }) = 0.05$.
  - But the conditional probability $P(\text{ Spam } \mid \text{ URGENT })$ is more useful for classifying new emails.

Bayes' rule gives a formula for finding $P(A \mid B)$ when $P(B \mid A)$ is known.

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}$$

Applying Bayes' rule to the penguins data.

Predicting penguin species using Bayes' rule

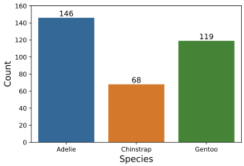
A = Adelie penguin

B = Body mass between 3750 g and 4000 g

Consider two events: A = the event a penguin is an Adelie penguin, and B = the event a penguin has a body mass between 3750 g and 4000 g.

Applying Bayes' rule to the penguins data.

Predicting penguin species using Bayes' rule



A = Adelie penguin

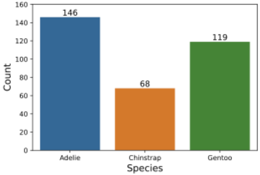
B = Body mass between 3750 g and 4000 g

$$P(A) = \frac{146}{146+68+119} = \frac{146}{333} = 0.4384$$

146 penguins are Adelie, 68 are Chinstrap, and 119 are Gentoo. The probability that a penguin selected at random from the sample is Adelie is $P(A) = \frac{146}{146+68+119} = 0.4384$.

Applying Bayes' rule to the penguins data.

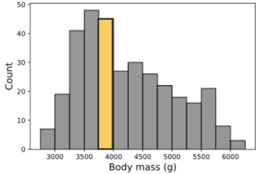
Predicting penguin species using Bayes' rule



A = Adelie penguin

B = Body mass between 3750 g and 4000 g

$$P(A) = \frac{146}{146+68+119} = \frac{146}{333} = 0.4384$$

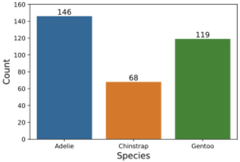


$$P(B) = \frac{45}{333} = 0.1351$$

45 penguins have a body mass between 3750 g and 4000 g, so $P(B) = \frac{45}{333} = 0.1351$.

Applying Bayes' rule to the penguins data.

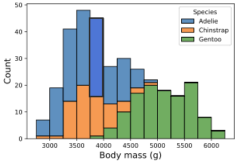
Predicting penguin species using Bayes' rule



A = Adelie penguin

B = Body mass between 3750 g and 4000 g

$$P(A) = \frac{146}{146+68+119} = \frac{146}{333} = 0.4384$$



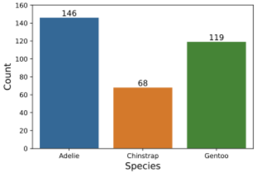
$$P(B) = \frac{45}{333} = 0.1351$$

$$P(B|A) = \frac{29}{146} = 0.1986$$

For the 146 Adelie penguins, 29 have a body mass between 3750 g and 4000 g. The probability an Adelie penguin has a body mass between 3750 g and 4000 g is $P(B|A) = \frac{29}{146} = 0.1986$.

Applying Bayes' rule to the penguins data.

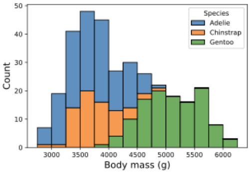
Predicting penguin species using Bayes' rule



A = Adelie penguin

B = Body mass between 3750 g and 4000 g

$$P(A) = \frac{146}{146+68+119} = \frac{146}{333} = 0.4384$$



$$P(B) = \frac{45}{333} = 0.1351$$

$$P(B|A) = \frac{29}{146} = 0.1986$$

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} = \frac{0.1986 \times 0.4384}{0.1351} = 0.6445$$

The probability a penguin with a body mass between 3750 g and 4000 g is Adelie is

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} = \frac{0.1986 \times 0.4384}{0.1351} = 0.6445.$$

Practice Question: Calculating conditional probabilities

Hospital staff would like to develop a diagnostic screening for heart disease based on high-density lipid (HDL) cholesterol. HDL cholesterol is considered the "good" cholesterol because HDL cholesterol helps remove other types of cholesterol from the bloodstream. A random sample of 500 patient records is collected. 39 patients in the sample have been diagnosed with heart disease. Let A denote the event a patient has heart disease and B denote the event a patient has low HDL cholesterol.

1) Calculate $P(A)$, the probability that a patient in the sample has heart disease.
   - ○ 0.039
   - ○ 0.078
   - ○ 0.922
2) 18% of patients in the sample had low HDL cholesterol. In probability notation, what does the 18% represent?
   - ○ $P(A|B)$
   - ○ $P(B|A)$
   - ○ $P(B)$
3) 27 of the patients with heart disease had low HDL cholesterol. Calculate $P(A|B)$, the probability that a patient with low HDL cholesterol has heart disease.
   - ○ 0.054
   - ○ 0.300
   - ○ 0.692

Practice Question: Calculating conditional probabilities

Hospital staff would like to develop a diagnostic screening for heart disease based on high-density lipid (HDL) cholesterol. HDL cholesterol is considered the "good" cholesterol because HDL cholesterol helps remove other types of cholesterol from the bloodstream. A random sample of 500 patient records is collected. 39 patients in the sample have been diagnosed with heart disease. Let A denote the event a patient has heart disease and B denote the event a patient has low HDL cholesterol.

1) Calculate $P(A)$, the probability that a patient in the sample has heart disease.
   - ○ 0.039
   - ◉ 0.078
   - ○ 0.922

Correct

39/500 = 0.078 is the probability a patient in the sample has heart disease.

2) 18% of patients in the sample had low HDL cholesterol. In probability notation, what does the 18% represent?
   - ○ $P(A|B)$
   - ○ $P(B|A)$
   - ◉ $P(B)$

Correct

B denotes the event a patient has low HDL cholesterol. Since no condition is specified, $P(B) = 0.18$.

3) 27 of the patients with heart disease had low HDL cholesterol. Calculate $P(A|B)$, the probability that a patient with low HDL cholesterol has heart disease.
   - ○ 0.054
   - ◉ 0.300
   - ○ 0.692

Correct

$0.300 = \frac{P(B|A)\times P(A)}{P(B)} = \frac{0.6923\times 0.078}{0.18}$ A patient with low HDL has a 30% probability of heart disease.

Naive Bayes Classifier

Naive Bayes classifier uses Bayes' rule to classify instances based on conditional probabilities. Let y_i denote class i of the output feature and x denote the input features.

- The **prior probability** represents the overall probability of class i, denoted $P(y_i)$.
- The **posterior probability** represents the probability of class i, given certain values of the input features x, denoted $P(y_i \mid x)$ Naive Bayes classifiers make predictions by calculating the posterior probabilities for all c classes.

By Bayes' rule,

$$P(y_i \mid x) = \frac{P(x \mid y_i) \times P(y_i)}{P(x)}$$

The class with the highest posterior probability becomes the predicted class for instance i

Classifying penguins using naive Bayes.

Classifying penguins using naive Bayes

x = Body mass between 3750 g and 4000 g

y_i = Species

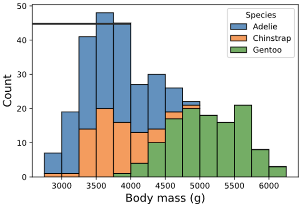
Let x denote the event a penguin has a body mass between 3750 g and 4000 g. y_i has three possible values: Adelie, Chinstrap, and Gentoo.

Classifying penguins using naive Bayes.

Classifying penguins using naive Bayes

x = Body mass between 3750 g and 4000 g

y_i = Species



$$P(x) = \frac{45}{333} = 0.1351$$

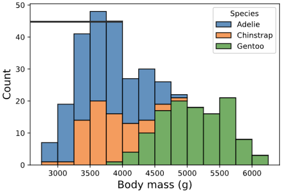
45 out of 333 penguins have a body mass between 3750 g and 4000 g. So, $P(x) = 0.1351$.

Classifying penguins using naive Bayes.

Classifying penguins using naive Bayes

x = Body mass between 3750 g and 4000 g

y_i = Species



$$P(x) = \frac{45}{333} = 0.1351$$

Prior probabilities

$$P(Adelie) = \frac{146}{333} = 0.4384$$

$$P(Chinstrap) = \frac{68}{333} = 0.2042$$

$$P(Gentoo) = \frac{119}{333} = 0.3574$$

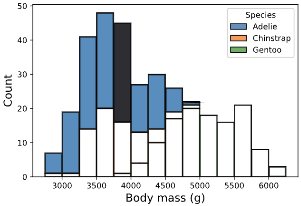
The prior probabilities are based on the sample proportions. 146 penguins in the dataset are Adelie, 68 penguins are Chinstrap, and 119 are Gentoo.

Classifying penguins using naive Bayes.

Classifying penguins using naive Bayes

x = Body mass between 3750 g and 4000 g

y_i = Species



$$P(x) = \frac{45}{333} = 0.1351$$

$$P(x|Adelie) = \frac{29}{146} = 0.1986$$

Prior probabilities

$$P(Adelie) = \frac{146}{333} = 0.4384$$

$$P(Chinstrap) = \frac{68}{333} = 0.2042$$

$$P(Gentoo) = \frac{119}{333} = 0.3574$$

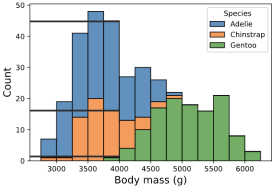
29 out of 146 Adelie penguins have a body mass between 3750 g and 4000 g, so $P(x|Adelie) = 29/146 = 0.1986$.

Classifying penguins using naive Bayes.

Classifying penguins using naive Bayes

x = Body mass between 3750 g and 4000 g

y_i = Species



$$P(x) = \frac{45}{333} = 0.1351$$

$$P(x|Adelie) = \frac{29}{146} = 0.1986$$

$$P(x|Chinstrap) = \frac{15}{68} = 0.2206$$

$$P(x|Gentoo) = \frac{1}{119} = 0.0084$$

Prior probabilities

$$P(Adelie) = \frac{146}{333} = 0.4384$$

$$P(Chinstrap) = \frac{68}{333} = 0.2042$$

$$P(Gentoo) = \frac{119}{333} = 0.3574$$

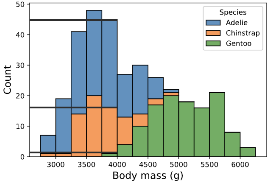
15 out of 68 Chinstrap penguins and 1 out of 119 Gentoo penguins have a body mass between 3750 g and 4000 g, so $P(x|Chinstrap) = 0.2206$ and $P(x|Gentoo) = 0.0084$.

Classifying penguins using naive Bayes.

Classifying penguins using naive Bayes

x = Body mass between 3750 g and 4000 g

y_i = Species



$$P(x) = \frac{45}{333} = 0.1351$$

$$P(x|Adelie) = \frac{29}{146} = 0.1986$$

$$P(x|Chinstrap) = \frac{15}{68} = 0.2206$$

$$P(x|Gentoo) = \frac{1}{119} = 0.0084$$

Prior probabilities

$$P(Adelie) = \frac{146}{333} = 0.4384$$

$$P(Chinstrap) = \frac{68}{333} = 0.2042$$

$$P(Gentoo) = \frac{119}{333} = 0.3574$$

Posterior probabilities

$$\boxed{P(Adelie|x) = \frac{0.1986 \times 0.4384}{0.1351} = 0.6445}$$

$$P(Chinstrap|x) = \frac{0.2042 \times 0.2206}{0.1351} = 0.3334$$

$$P(Gentoo|x) = \frac{0.3574 \times 0.0084}{0.1351} = 0.0222$$

Applying Bayes' rule results in the posterior probabilities. $P(Adelie|x)$ is the highest posterior probability, so the predicted class is Adelie.

Classifying penguins using naive Bayes.

Hospital staff would like to develop a diagnostic screening for heart disease based on HDL cholesterol. A random sample of 500 patient records is collected. Patients are categorized as having low HDL or healthy HDL. Let x be the HDL level, and y_i be 1 if a patient has heart disease, and 0 if a patient does not have heart disease.

Diagnosis	Low HDL	Healthy HDL	Total
Heart disease	27	12	39
No heart disease	63	398	461
Total	90	410	500

1) Calculate $P(y_i = 1 | x = \text{low})$.
   - 0.300
   - 0.429
   - 0.692

2) Calculate $P(y_i = 1 | x = \text{healthy})$.
   - 0.029
   - 0.095
   - 0.308

3) Calculate $P(y_i = 0 | x = \text{healthy})$.
   - 0.700
   - 0.796
   - 0.971

4) Using naive Bayes, how should a patient with healthy HDL be classified?
   - Heart disease
   - No heart disease

Classifying penguins using naive Bayes.

Hospital staff would like to develop a diagnostic screening for heart disease based on HDL cholesterol. A random sample of 500 patient records is collected. Patients are categorized as having low HDL or healthy HDL. Let x be the HDL level, and y_i be 1 if a patient has heart disease, and 0 if a patient does not have heart disease.

Diagnosis	Low HDL	Healthy HDL	Total
Heart disease	27	12	39
No heart disease	63	398	461
Total	90	410	500

1) Calculate $P(y_i = 1 | x = \text{low})$.

- (selected) 0.300
- 0.429
- 0.692

Correct

$P(y_i = 1 | x = \text{low}) = \frac{27}{90} = 0.300$. 30% of patients with low HDL cholesterol have heart disease.

2) Calculate $P(y_i = 1 | x = \text{healthy})$.

- (selected) 0.029
- 0.095
- 0.308

Correct

$P(y_i = 1 | x = \text{healthy}) = \frac{12}{410} = 0.029$. 2.9% of patients with healthy HDL levels have heart disease.

3) Calculate $P(y_i = 0 | x = \text{healthy})$.

- 0.700
- 0.796
- (selected) 0.971

Correct

Whether or not a patient has heart disease is a binary outcome. So,
$P(y_i = 0 | x = \text{healthy}) = 1 - P(y_i = 1 | x = \text{healthy})$
$= 1 - 0.029 = 0.971$.

4) Using naive Bayes, how should a patient with healthy HDL be classified?

- Heart disease
- (selected) No heart disease

Correct

$P(y_i = 0 | x = \text{healthy}) > P(y_i = 1 | x = \text{healthy})$
So, a patient with healthy HDL is classified as $y_i = 0$.

Naive Bayes Assumptions

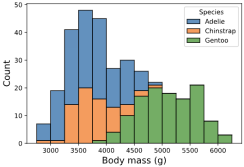
The name “naive Bayes" refers to a set of assumptions built into the naive Bayes classifier. Naive Bayes classification assumes:

1. All input features are independent or uncorrelated.
2. All input features are equally important.

But in reality, the naive Bayes assumptions are rarely satisfied. The naive Bayes assumptions can be evaluated by exploring the input features and the data context.

Naive Bayes Assumptions

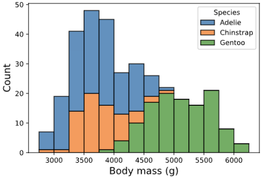
Input: Body mass



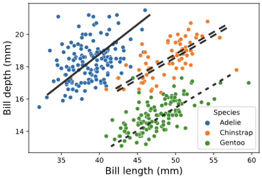
The naive Bayes assumptions are always met for a single input feature, since no relationship to other input features can exist.

Naive Bayes Assumptions

Input: Body mass



Inputs: Bill length and bill depth

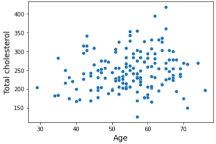


A relationship exists between bill length and bill depth: Penguins in each species with longer bills have deeper bills. The naive Bayes assumption of independence is not met.

Practice Problem: Evaluating the naive Bayes assumptions.

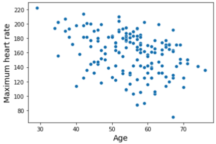
Hospital staff would like to develop a diagnostic screening for heart disease based on HDL cholesterol. But other features may also be good inputs, such as age, total cholesterol, and maximum heart rate, during a cardiac stress test.

1) For age and total cholesterol, the naive Bayes independence assumption is _____.



- ○ met
- ○ violated

2) For age and maximum heart rate, the naive Bayes independence assumption is _____.



- ○ met
- ○ violated

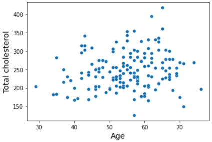
3) Who should a data scientist consult to determine which features are clinically important?

- ○ A cardiologist
- ○ Another data scientist
- ○ A patient with heart disease

Practice Problem: Evaluating the naive Bayes assumptions.

Hospital staff would like to develop a diagnostic screening for heart disease based on HDL cholesterol. But other features may also be good inputs, such as age, total cholesterol, and maximum heart rate, during a cardiac stress test.

1) For age and total cholesterol, the naive Bayes independence assumption is _____.

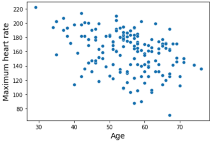


- (selected) met
- violated

Correct

As age increases, total cholesterol does not change. No relationship exists between age and total cholesterol, so the naive Bayes independence assumption is met.

2) For age and maximum heart rate, the naive Bayes independence assumption is _____.



- met
- (selected) violated

Correct

As age increases, maximum heart rate decreases. A negative relationship exists between age and maximum heart rate, so the naive Bayes independence assumption is not met.

3) Who should a data scientist consult to determine which features are clinically important?

- (selected) A cardiologist
- Another data scientist
- A patient with heart disease

Correct

Cardiologists routinely work with heart disease patients and are experts in how the heart works. Cardiologists can provide guidance on which features are clinically important.

Gaussian naive Bayes

For categorical or discrete input features, sample probabilities can be calculated for each individual value of x. For numerical input features, continuous probability distributions are used instead of sample probabilities. A **continuous probability distribution** is a mathematical function that describes the probability that a certain value of a random variable occurs.

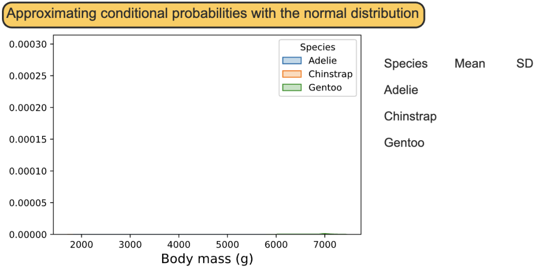
The most common choice for numerical input features is the Gaussian, or normal, distribution. The **normal distribution**, denoted normal (μ, σ), is a symmetric, bell-shaped distribution with two parameters: the mean, μ, and the standard deviation σ. The normal distribution provides a good approximation for many input features.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left((x-\mu)^2/2\sigma^2\right)$$

Gaussian naive Bayes

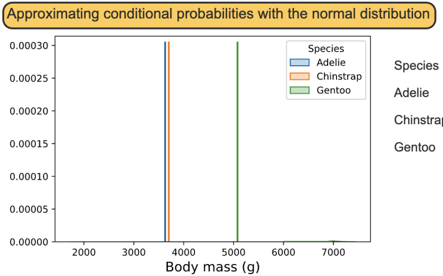
Gaussian naive Bayes uses the normal distribution as an approximation to the conditional probability $P(x \mid y_i)$. One normal distribution is fitted to each class and used to calculate the posterior probabilities.

Approximating Conditional Probabilities with the Normal Distribution



The Gaussian, or normal, distribution has two parameters: μ and σ. In practice, μ and σ are estimated from the sample data.

Approximating Conditional Probabilities with the Normal Distribution

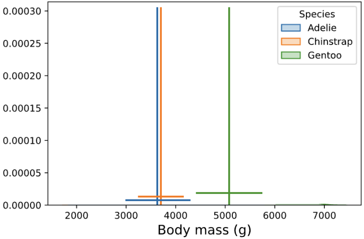


Species	Mean	SD
Adelie	3706.2	
Chinstrap	3733.1	
Gentoo	5092.4	

μ sets the normal distribution's mean, or center. Gentoo penguins have the highest mean body mass, and Adelie penguins have the lowest.

Approximating Conditional Probabilities with the Normal Distribution

Approximating conditional probabilities with the normal distribution

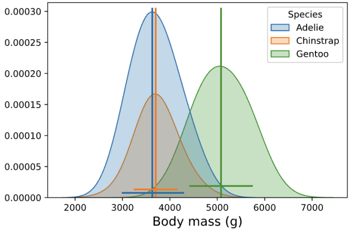


Species	Mean	SD
Adelie	3706.2	458.6
Chinstrap	3733.1	384.3
Gentoo	5092.4	501.5

σ sets the normal distribution's standard deviation, or spread. Chinstrap penguins have the lowest spread, and Gentoo penguins have the highest.

Approximating Conditional Probabilities with the Normal Distribution

Approximating conditional probabilities with the normal distribution



Species	Mean	SD
Adelie	3706.2	458.6
Chinstrap	3733.1	384.3
Gentoo	5092.4	501.5

Three normal distributions are plotted: one for each species. Each distribution represents the conditional probability $P(x|y_i)$.

Practice Problem: Normal approximation for heart disease screening.

A follow-up study selected a sample of 100 patients with heart disease and 100 patients without heart disease to participate in an experiment. Patients were asked to take a cardiac stress test, and the maximum heart rate was recorded for each patient. The mean and standard deviation of heart rate for each group is in the table below.

Diagnosis	Mean	SD
Heart disease	177.5	20.05
No heart disease	140.3	23.44
All patients	158.9	28.69

1) The normal distribution will be used to approximate the distribution of $P(x|y_i)$ for each ____.
   - group of patients
   - individual patient
   - maximum heart rate

2) For patients with heart disease, the normal distribution's mean is μ = ____.
   - 140.3
   - 158.9
   - 177.5

3) For patients without heart disease, the normal distribution's standard deviation is σ = ____.
   - 20.05
   - 23.44
   - 28.69

4) The normal curve for heart disease patients at x = 200 will be ____ the normal curve patients without heart disease.
   - higher than
   - lower than
   - equal to

Practice Problem: Normal approximation for heart disease screening.

A follow-up study selected a sample of 100 patients with heart disease and 100 patients without heart disease to participate in an experiment. Patients were asked to take a cardiac stress test, and the maximum heart rate was recorded for each patient. The mean and standard deviation of heart rate for each group is in the table below.

Diagnosis	Mean	SD
Heart disease	177.5	20.05
No heart disease	140.3	23.44
All patients	158.9	28.69

1) The normal distribution will be used to approximate the distribution of $P(x|y_i)$ for each ____.

- (selected) group of patients
- individual patient
- maximum heart rate

Correct

The experiment uses two groups: patients with heart disease and patients without heart disease. The normal distribution will approximate the distribution of maximum heart rate (x) for each group of patients (y_i).

2) For patients with heart disease, the normal distribution's mean is μ = ____.

- 140.3
- 158.9
- (selected) 177.5

Correct

The mean maximum heart rate for patients with heart disease is 177.5. Compared to the mean for patients without heart disease (140.3), patients with heart disease have a faster heart rate.

3) For patients without heart disease, the normal distribution's standard deviation is σ = ____.

- 20.05
- (selected) 23.44
- 28.69

Correct

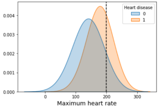
The standard deviation of maximum heart rate for patients with heart disease is 23.44. Maximum heart rate is more spread out for patients without heart disease.

4) The normal curve for heart disease patients at x = 200 will be ____ the normal curve patients without heart disease.

- (selected) higher than
- lower than
- equal to

Correct

The mean maximum heart rate for heart disease patients is closer to x = 200, so the normal curve is higher for heart disease patients than for patients without heart disease.



Bayesian Priorsi

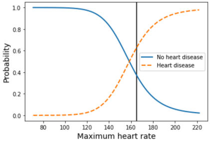
Bayesian models, including naive Bayes classifiers, incorporate prior assumptions about the probability a given event occurs in the model's predictions.

- Ex: By default, most implementations of naive Bayes classification use the sample probabilities of each class, $P(y_i)$, as the prior probabilities.
- But the prior probabilities may be adjusted based on outside information. Adjusting the prior probabilities may have an impact on the model's predictions and performance.

Effect of prior probabilities on predictions.

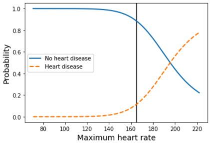
Two Gaussian naive Bayes models were fitted to the cardiac stress test experiment. The probability curves for each model are shown below.

1) Model 1 assumed a "uniform prior": $P(y_i = 1) = P(y_i = 0) = 0.5$. Which group had a higher posterior probability at maximum heart rate = 165 beats per minute?



- Heart disease
- No heart disease

2) In the US population, about 7.2% of adults have heart disease. Model 2 assumed that $P(y_i = 1) = 0.072$ and $P(y_i = 0) = 0.928$. Which group had a higher posterior probability at maximum heart rate = 165 beats per minute?



- Heart disease
- No heart disease

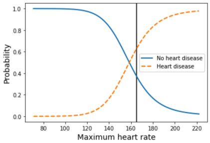
3) Which model is more reflective of reality?

- Model 1
- Model 2

Effect of prior probabilities on predictions.

Two Gaussian naive Bayes models were fitted to the cardiac stress test experiment. The probability curves for each model are shown below.

1) Model 1 assumed a "uniform prior": $P(y_i = 1) = P(y_i = 0) = 0.5$. Which group had a higher posterior probability at maximum heart rate = 165 beats per minute?

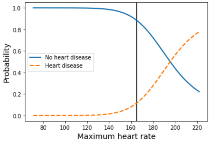


- (selected) Heart disease
- No heart disease

Correct

$P(y_i = 1|x = 165) = 0.627$ and $P(y_i = 0|x = 165) = 0.373$. According to Model 1, a patient with maximum heart rate = 165 beats per minute is more likely to have heart disease.

2) In the US population, about 7.2% of adults have heart disease. Model 2 assumed that $P(y_i = 1) = 0.072$ and $P(y_i = 0) = 0.928$. Which group had a higher posterior probability at maximum heart rate = 165 beats per minute?



- Heart disease
- (selected) No heart disease

Correct

$P(y_i = 1|x = 165) = 0.116$ and $P(y_i = 0|x = 165) = 0.884$. According to Model 2, a patient with maximum heart rate = 165 beats per minute is more likely to not have heart disease.

3) Which model is more reflective of reality?

- Model 1
- (selected) Model 2

Correct

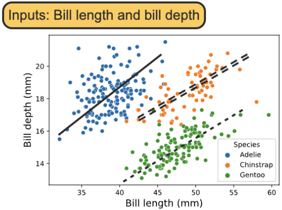
Even though the cardiac stress test experiment had equal proportions, in reality, the proportion of people with heart disease is much less than 50%. Bayesian techniques like naive Bayes allow data scientists to work outside information, like the population rate of heart disease, into an analysis.

Advantages and disadvantages

Naive Bayes' predictions are fast to compute, since predictions are based on conditional probabilities.

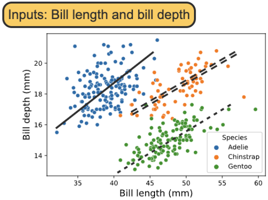
- For large datasets that require fast predictions, the computational advantages make naive Bayes a good choice. But the naive Bayes assumptions are often unrealistic.
- A tradeoff exists between computational ease and theoretical requirements. If the predictions are fast and accurate, naive Bayes may still be useful despite violated assumptions.

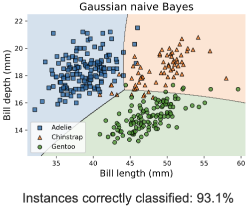
Naive Bayes Assumptions.



A relationship exists between bill length and bill depth: Most penguins with longer bills also have deeper bills. The naive Bayes assumption of independence is not met.

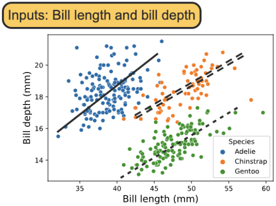
Naive Bayes Assumptions.

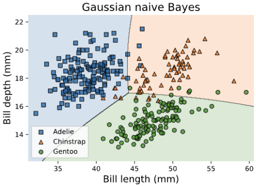




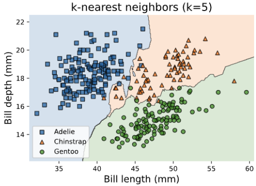
Despite the violated assumptions, predictions from Gaussian naive Bayes are accurate. 93.1% of instances are correctly classified.

Naive Bayes Assumptions.





Instances correctly classified: 93.1%



Instances correctly classified: 97.3%

k-nearest neighbors with $k = 5$ correctly classifies 97.3% of instances. k-nearest neighbors is more accurate, but also more computationally complex.